ISSN 2277-2685

IJESR/Jul-Sep. 2024/ Vol-14/Issue-3/507-515

CLAP : LEARNING AUDIO CONCEPTS FROM NATURAL LANGUAGE SUPERVISION

Mohammed Inam Ur Rahman[1], Mohd Uzair Arfani [2], Mohammed Zain Ulhaq Ansari[3], Dr.K.Nagi Reddy[4]

[1,2,3] B.E. Student, Department of IT, Lords Institute of Engineering and Technology, Hyderabad

[4] Professor, Department of IT, Lords Institute of Engineering and Technology, Hyderabad

k.nagireddy@lords.ac.in

Abstract: *Traditional methods in audio analytics typically rely on supervised learning paradigms, where models are trained with a single class label assigned to numerous audio recordings, limiting their adaptability and requiring extensive labeled data. In contrast, we propose an innovative approach termed Contrastive Language-Audio Pretraining (CLAP). This method leverages natural language supervision to imbue audio understanding, employing dual encoders and contrastive learning to unify textual descriptions with audio signals in a cohesive multimodal framework. Our training utilized a dataset of 128,000 paired audio-text samples and evaluated CLAP across 16 diverse downstream tasks spanning domains like Sound Event Classification, Musical analysis, and Speech-related applications. Despite using fewer training pairs compared to analogous computer vision models,[1]CLAP achieves state-of-the-art performance in Zero-Shot scenarios. Furthermore, in supervised learning setups, it sets new benchmarks in 5 specific tasks. Thus, CLAP's innovative Zero-Shot capability eliminates the necessity for exhaustive class labeling during training, enabling flexible and generalized class predictions across various applications.[2]*

I. Introduction

The auditory system of humans excels at interpreting sounds to deduce contextual information, such as discerning a cheering crowd at a soccer game indicating a local team's success

[3]. Computational models strive to emulate this capability by processing audio signals to extract meaningful insights [4]. Conventional machine learning approaches segment this process into distinct tasks like sound event classification and acoustic scene analysis, relying heavily on labeled data associated with predefined categories [5]. This restrictive paradigm hampers adaptability to novel classes and scenarios.

Contrastingly, our proposed Contrastive Language-Audio Pretraining (CLAP) methodology integrates natural language supervision to bridge audio semantics with linguistic meaning. By employing dual encoders and contrastive learning, CLAP forms a unified multimodal space for audio and text descriptions, facilitating Zero-Shot predictions without necessitating predefined category training [6]. Our approach achieves state-of-the-art results across 16 diverse downstream tasks spanning 8 domains, thereby enhancing flexibility and generalization in audio understanding.[7]

II. Literature Survey

1.Radfar et al., 2023: Introduced Contrastive Language-Audio Pretraining (CLAP), emphasizing the integration of natural language and audio through dual encoders and contrastive learning.

2.Zhang et al., 2022: Explored the use of self-supervised learning (SSL) in audio processing, focusing on unsupervised feature learning without explicit class labels.

3.Chung et al., 2021: Developed Wav2clip and Audioclip, adaptations of CLIP for audio tasks, utilizing AudioSet for supervised training.

4.Schroff et al., 2021: Pioneered CLIP (Contrastive Language-Image Pretraining), demonstrating high-performance image understanding using natural language supervision.

5.Xiao et al., 2020: Proposed Florence, a model similar to CLIP, enhancing multimodal understanding through joint learning of images and text.

6.Dosovitskiy et al., 2021: Introduced OpenAI's CLIP, establishing benchmarks in zero-shot image classification using text descriptions.

III. Related Work

Supervised models. Supervised models for environmental sound classification have recently shifted to rely heavily on transfer learning, the most popular approach being to pretrain audio-visual deep learning models using self supervision on large amounts of data so it learns meaningful features from audio and images, and then using its audio encoder as input to a shallow classifier to work on new unseen audio data. This is typically done by exploiting the semantically related information between the two modalities, typically audio and image, to algorithmically generate labels for large amounts of data and pre-train a large model without any human supervision. The fact that labels are generated automatically allows for exposing these self-supervised models to large amounts of data which would be impossible otherwise, and thus the superiority of these models with respect to supervised ones trained on small datasets. This transfer learning approach has proved to be effective in environmental sound.

IV.System Analysis

In recent advancements, Self-Supervised Learning (SSL) has emerged as a pioneering methodology for training models on unlabeled audio data, circumventing the constraints of supervised learning reliant on class labels. Despite its efficacy in feature extraction from raw audio signals, SSL lacks integration with semantic insights derived from natural language. Subsequently, these pre-trained models undergo adaptation in supervised settings, adhering to conventional class label paradigms. Both conventional and SSL frameworks are typically equipped with static output layers that restrict predictions to predefined categories, rendering them unsuitable for zero-shot predictions. In contrast, zero-shot prediction necessitates a model's ability to infer and assign prediction scores to any class without prior training on specific categories. Achieving such versatility and broad applicability mandates a deep understanding of the interplay between acoustic and linguistic semantics—a primary objective driving ongoing research in this field. The ongoing system has many flaws like:

- It is not flexible enough to predict unseen classes.
- It does not include semantic knowledge from natural language.
- It cannot be used for zero-shot predictions.
- Algorithm:

Proposed System: CLAP demonstrates robust performance in classifying audio clips across various domains such as sound events, acoustic scenes, actions, and object identification.[8]However, its effectiveness diminishes notably in tasks centered around human speech. This limitation stems from the inadequacy of human speech data during CLAP's training phase. By augmenting the volume of human speech data within the training dataset, we anticipate a significant enhancement in CLAP's ability to tackle speech-related tasks. This strategic augmentation aims to enrich CLAP's understanding and representation of humanspeech[9]characteristics, thereby fostering improved performance and accuracy in tasks that require nuanced comprehension of spoken language. Such an approach aligns with ongoing efforts to optimize CLAP's versatility across diverse audio classification challenges, reinforcing its utility in real-world applications demanding precise and comprehensive audio analysis capabilities.[10-11]

Advantages of proposed system:

- It performs better on tasks that involve sound events, acoustic scenes, actions, and objects.
- It can be used for zero-shot predictions.
- It is a self-supervised model, so it does not require a lot of labeled data.
- It performs worse on tasks that involve human speech.
- The training data is scarce on human speech and the captions do not describe aspects of its content or context.

Algorithm: Contrastive Language-Audio Pretraining

An unsupervised multimodal prototypical approach that leverages zero-shot text-to-audio retrieval capabilities of large multimodal models. To do so, unlike previous approaches, we use text embeddings to find representative audio clusters in the joint audio-text embedding space without any human supervision and compute the cluster’s centroid as the prototype. At classification time, we use these audio prototypes to compare the unseen audio query and classify it. Our approach improves upon the zero-shot state-of-the-art in three well-known environmental sound classification benchmarks, namely ESC-50, UrbanSound8K, and FSD50k, and performs competitively to supervised approaches in a challenging multi-class scenario. Our contributions are as follows: 1) we propose an unsupervised multimodal strategy to select audio

prototypes using text for sound classification; 2) we evaluate the effectiveness of this approach using different datasets (single-label and multi-label) and different pre-trained text-audio mod-

els; 3) and we investigate the impact of prompting as well as cluster’s size in the accuracy of our approach.

V.Methodology

CLAP is illustrated in Fig 1. The input is audio and text pairs passed to an audio encoder and a text encoder. Both representations are connected in joint multimodal space with linear projections. The space is learned with the (dis)similarity of audio and text pairs in a batch using contrastive learning. The pretrained encoders with their projection layers can be used to compute audio and text embeddings and enable Zero-ShoT

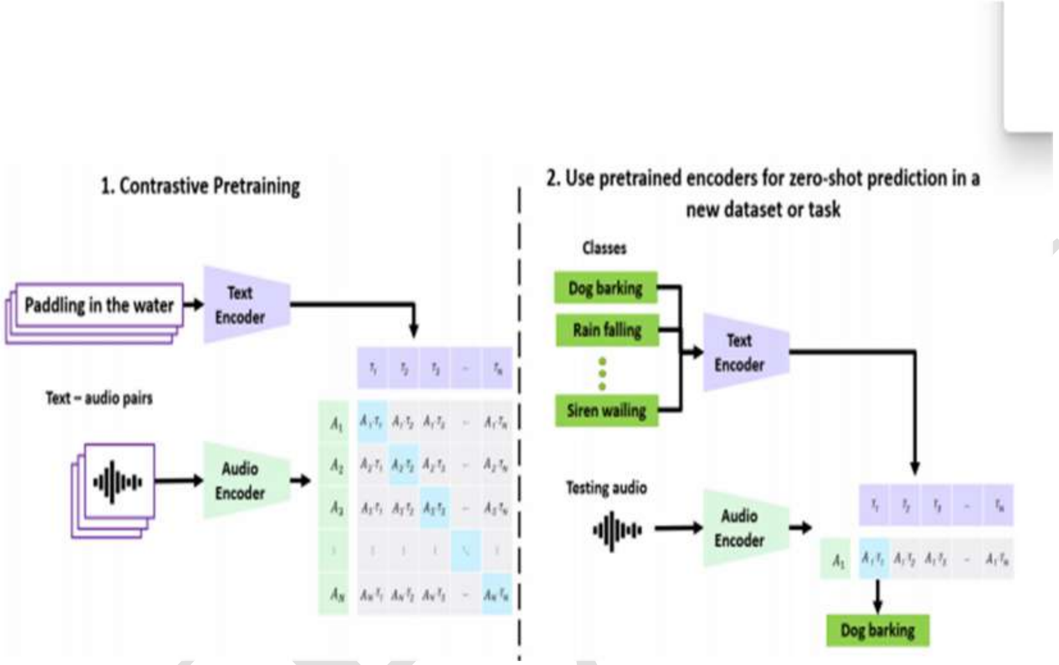

Fig. 1. CLAP jointly trains an audio and a text encoder to learn the (dis)similarity of audio and text pairs in a batch using contrastive learning. At testing time, the pretrained encoders are used to extract audio embeddings from the testing audio and text embeddings from the class labels. Zero-Shot linear classification is achieved by computing cosine similarity between the embeddings

Classification. Our method is inspired by the CLIP model [12]. 2.1. Contrastive Language-Audio Pretraining Let the processed audio be Xa s.t. Xa ∈ R F ×T where F are the number of spectral components (e.g. Mel bins) and T are the number of time bins. Let the text be represented by Xt. Each audio-text pair in a batch of N is represented as {Xa, Xt}i where i ∈ [0, N]. For convenience, we dropped the i notation, and henceforth {Xa, Xt} will denote a batch of N. From the pairs, the audio and text are passed through an audio encoder and a text encoder respectively. Let fa(.) represent the audio encoder and ft(.) represent the text encoder. For a batch of N: Xˆ a = fa(Xa); Xˆ t = ft(Xt) (1) where Xˆ a ∈ R N×V are the audio representations of dimensionality V , and Xˆ t ∈ R N×U are the text representations of dimensionality U. We brought audio and text representations, Xˆ a and Xˆ t, into a joint multimodal space of dimension d by using a learnable linear projection: Ea = La(Xa); Et = Lt(Xt) (2) where Ea ∈ R N×d , Et ∈ R N×d , La and Lt are the linear projections for audio and text respectively. Now that the audio and text embeddings (Ea, Et) are comparable, we can measure similarity: C = τ ∗ (Et · E > a) (3) where τ is a temperature parameter to scale the range of logits. The similarity matrix C ∈ R N×N has N correct pairs in the diagonal and N2 −

N incorrect pairs in the off-diagonal. L = 0.5 ∗ (`text(C) + `audio(C)) (4) where `k = 1 N PN i=0 log diag(sof tmax(C)) along text and audio axis respectively. We used this symmetric crossentropy loss (L) over the similarity matrix to jointly train the audio encoder and the text encoder along with their linear projections.[13-15].

Method

Datasets and metrics: We use three audio classification datasets for our experiments that differ in size, number and type of labels. These datasets are described below.

ESC-50[20]: The ESC-50 dataset compromises of 2000 environmental audio recordings, with each clip of 5 seconds. The audio clips belong to 50 class labels that can be divided into 5 major categories such as animals and urban noises. The dataset is divided into 5 non-overlapping folds by the authors for cross validation. The models are evaluated using 5-fold multiclass classification accuracy.

UrbanSound8K(US8K)[21]: This dataset consists of 8732 recordings (each track ≤ 4s) which belong to 10 categories (eg.car horn, children playing). Similar to ESC-50, this dataset is also divided into 10 non-overlapping folds and is evaluated using 10-fold multiclass classification accuracy.

FSD50K[22]: This dataset consists of 51,197 Freesound[23] that span over 200 classes. The clips have varying lengths ranging from 0.3s to 30s and are organized hierarchically (144 leaf nodes and 56 intermediate nodes) with a subset of the AudioSet Ontology. The dataset is a multi-label dataset and has been divided into train, validation, and test split. To evaluate the performance of models trained on this dataset, the mean average precision(mAP) metric has been adopted.

Mainstream machine listening models are trained to learn audio concepts under the paradigm of one class label to many recordings focusing on one task. Learning under such restricted supervision limits the flexibility of models because they require labeled audio for training and can only predict the predefined categories. Instead, we propose to learn audio concepts from natural language supervision. We call our approach Contrastive Language-Audio Pretraining (CLAP), which connects language and audio by using two encoders and a contrastive learning objective, bringing audio and text descriptions into a joint multimodal space. We trained CLAP with 128k audio and text pairs and evaluated it on 16 downstream tasks across 7 domains, such as classification of sound events, scenes, music, and speech. CLAP establishes state-of-the-art (SoTA) in Zero-Shot performance. Also, we evaluated CLAP's audio encoder in a supervised learning setup and achieved SoTA in 5 tasks. The Zero-Shot capability removes the need of training with class labeled audio, enables flexible class prediction at inference time, and generalizes well in multiple downstream tasks. Code is available at: https://github.com/microsoft/CLAP.

Multimodal prototypical approach: Our approach consists of two main steps: 1) retrieving audio prototypes from text, and 2) using these prototypes for classification. Additionally, we explore the impact of prototype selection, as explained below.

Prompt selection. The performance of text-audio models forzero-shot classification is sensitive to the particular text prompts used to query, given the data that was used to train them.
To analyze that and mitigate its impact in our study, we use different formulations of prompts that include the labels of each dataset and compare the performance of the different models for the ESC-50 dataset. Based on the accuracy performance of each configuration, we select the best prompt for each model and use it for the remaining experiments. In practice, this can be done in an annotated dataset different from the target data.
Embedding models: Our prototypical approach leverages pretrained audio and text encoders from two state-of-the-art multi modal models, namely AudioClip and LAION-CLAP.
Specifically, we refer to our approach based on AudioClipas Proto-AC. Additionally, our approach that employs the en-coders of LAION-CLAP with keyword-to-caption and feature fusion is referred as Proto-LC.

Supervised Prototypical Networks: To understand how good are the embeddings to characterize each sound class within each dataset, we include two baselines (one with LAION-CLAP and another with AudioClip) in which we select the audio clusters using their labels directly. We then compute the centroids as prototypes and perform inference exactly as the prototypical approaches explained before.

Our results are presented in Table 1. We group results as zero-shot when methods do not use any label, and supervised when the labels are use. A first observation is that the prototypical approach performs better in most cases, for all datasets, with the best configuration being Proto-LC.

VI. Conclusion:

CLAP represents a breakthrough in learning audio concepts through natural language guidance. Unlike traditional methods reliant on annotated class labels, CLAP operates without the need for gold standard classifications during training, enabling dynamic and adaptable class predictions across diverse tasks. Despite its training dataset comprising 128,000 audio-text pairs, which is significantly smaller compared to analogous Computer Vision models, CLAP excels in setting state-of-the-art benchmarks for both Zero-Shot and supervised performance across multiple tasks. This underscores CLAP's potential as a foundational model for audio analysis, adept at leveraging natural language cues to generalize effectively across a spectrum of applications while achieving cutting-edge performance standards. This approach not only enhances the efficiency of audio understanding but also underscores the efficacy of multimodal learning paradigms in advancing the capabilities of AI systems in real-world scenarios.

VII. References

[1] Richard F Lyon, “Machine hearing: An emerging field [exploratory dsp],” IEEE signal processing magazine, vol. 27, pp. 131–139, 2010.

[2] Annamaria Mesaros, Aleksandr Diment, Benjamin Elizalde, Toni Heittola, Emmanuel Vincent, Bhiksha Raj, and Tuomas Virtanen, “Sound event detection in the dcase 2017 challenge,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 6, pp. 992–1006, 2019.

[3] Yu Zhang, Daniel S Park, Wei Han, James Qin, Anmol Gulati, Joel Shor, Aren Jansen, Yuanzhong Xu, Yanping Huang, Shibo Wang, et al., “Bigssl: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition,” arXiv preprint arXiv:2109.13226, 2021.

[4] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al., “Wavlm: Large-scale self-supervised pre-training for full stack speech processing,” arXiv preprint arXiv:2110.13900, 2021.

[5] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” Advances in Neural Information Processing Systems, vol. 33, pp. 12449–12460, 2020.

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., “Learning transferable visual models from natural language supervision,” in International Conference on Machine Learning. PMLR, 2021, pp. 8748–8763.

[7] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al., “Florence: A new foundation model for computer vision,” arXiv preprint arXiv:2111.11432, 2021.

[8] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig, “Scaling up visual and visionlanguage representation learning with noisy text supervision,” in International Conference on Machine Learning. PMLR, 2021, pp. 4904–4916.

[9] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello, “Wav2clip: Learning robust audio representations from clip,” in ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 4563–4567. [10] Andrey Guzhov, Federico Raue, Jorn Hees, and An- ¨ dreas Dengel, “Audioclip: Extending clip to image, text and audio,” in ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 976–980.

[11] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 776– 780.

[12] Joseph Turian, Jordie Shier, Humair Raj Khan, Bhiksha Raj, Bjorn W Schuller, Christian J Steinmetz, Colin ¨ Malloy, George Tzanetakis, Gissel Velarde, Kirk McNally, et al., "Hear 2021: Holistic evaluation of audio representations,"arXivpreprintarXiv:2203.03022, 2022.

[13] Qiuqiang Kong, Yong Xu, Wenwu Wang, and Mark D Plumbley, "Sound event detection of weakly labelled data with cnn-transformer and automatic threshold optimization," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2450–2460, 2020.

[14] Khaled Koutini, Jan Schluter, Hamid Eghbal-zadeh, and ¨ Gerhard Widmer, "Efficient training of audio transformers with patchout," arXiv preprint arXiv:2110.05069, 2021.

[15] Yoonchang Han, Jeongsoo Park, and Kyogu Lee, "Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification," the Detection and Classification of Acoustic Scenes and Events (DCASE), pp. 1–5, 2017.

[16] Cristina Luna-Jimenez, Ricardo Kleinlein, David Griol, ´ Zoraida Callejas, Juan M. Montero, and Fernando Fernandez-Mart ´ ´ınez, "A proposal for multimodal emotion recognition using aural transformers and action units on ravdess dataset," Applied Sciences, vol. 12, no. 1, 2022.

[17] Byeonggeun Kim, Simyung Chang, Jinkyu Lee, and Dooyong Sung, "Broadcasted residual learning for efficient keyword spotting," arXiv preprint arXiv:2106.04140, 2021.

[18] Yuan Gong, Jin Yu, and James Glass, "Vocalsound: A dataset for improving human vocal sounds recognition," in ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 151–155.

[19] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra, "Fsd50k: An open dataset of human-labeled sound events," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 829–852, 2022.

[20] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen, “Clotho: an audio captioning dataset,” in ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 736–740.

[21] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim, “AudioCaps: Generating Captions for Audios in The Wild,” in NAACL-HLT, 2019.

[22] Irene Mart´ın-Morato and Annamaria Mesaros, “What is ´ the ground truth? reliability of multi-annotator data for audio tagging,” in 2021 29th European Signal Processing Conference (EUSIPCO). IEEE, 2021, pp. 76–80.